

# Validity and the foundations of statistical inference

Ryan Martin

Department of Mathematics, Statistics, and Computer Science  
University of Illinois at Chicago  
`rgmartin@uic.edu`

Chuanhai Liu

Department of Statistics  
Purdue University  
`chuanhai@purdue.edu`

July 19, 2016

## Abstract

In this paper, we argue that the primary goal of the foundations of statistics is to provide data analysts with a set of guiding principles that are guaranteed to lead to valid statistical inference. This leads to two new questions: “what is valid statistical inference?” and “do existing methods achieve this?” Towards answering these questions, this paper makes three contributions. First, we express statistical inference as a process of converting observations into degrees of belief, and we give a clear mathematical definition of what it means for statistical inference to be valid. Second, we evaluate existing approaches Bayesian and frequentist approaches relative to this definition and conclude that, in general, these fail to provide valid statistical inference. This motivates a new way of thinking, and our third contribution is a demonstration that the *inferential model* framework meets the proposed criteria for valid and prior-free statistical inference, thereby solving perhaps the most important unsolved problem in statistics.

*Keywords and phrases:* Bayes; belief; fiducial; inferential model; random set.

## 1 Introduction

At a basic level, statistics is concerned with the collection and analysis of data with the goal of testing existing theories and creating new ones. This is the cornerstone of the scientific method and, therefore, a solid foundation of statistics is fundamental to the advancement of science. The process of converting data into some kind of summary relevant to the scientific question is generally referred to as “statistical inference,” a term widely used in literature, including introductory- and advanced-level textbooks. However, as far as we know, the available literature has no agreed-upon definition of statistical

inference. A subject that lacks a proper definition of its main objective will inevitably face difficulties. For years, the use of p-values for significance testing has been criticized (e.g., Fidler et al. 2004); in fact, the journal *Basic and Applied Social Psychology* has recently banned the use of p-values (Trafimowa and Marks 2015). More important are the growing concerns about irreproducibility of experiments that have previously shown statistically significant results; see, e.g., Nuzzo (2014), the collection of reports published by *Nature*<sup>1</sup> on “Challenges in Irreproducible Research,” and the recent statement<sup>2</sup> issued by American Statistical Association. These existing challenges highlight the need for better understanding of our foundations. Barnard (1985) writes:

*I shall be concerned with the foundations of the subject. But in case it should be thought that this means I am not here strongly concerned with practical applications, let me say that confusion about the foundations of the subject is responsible, in my opinion, for much of the misuse of statistics that one meets in fields of application. . .*

Beyond these existing difficulties, of potentially greater concern is that there will be new challenges to overcome as we venture into so-called “big-data” problems. Indeed, as data sets get larger and models become more complex, the need for solid foundations becomes more pressing. Reid and Cox (2015) write:

*...for any analysis that hopes to shed light on the structure of the problem, modeling and calibrated inferences ... seem essential.*

On this point, we could not agree more. Unfortunately, existing foundational work is unclear about what is meant by “calibrated inference” or how it can be achieved in general. To us, the primary goal of the foundations of statistics is to provide a set of guiding principles that, if followed, will guarantee validity of the resulting inference. Our motivation for writing this paper is to be clear about what is meant by valid inference and to provide the necessary principles to help data analysts achieve validity.

Our first contribution, in Section 2, is to present a definition of statistical inference, formalizing the notion of converting information in data into beliefs about the unknown parameter. Our departure from standard approaches is to put primary emphasis on the assessment of uncertainty about the parameter of interest, rather than on decision procedures with certain sampling distribution properties. Of course, an important concern is that the data analyst’s beliefs should be meaningful for scientific applications, and this requires a certain calibration property. We define what it means for statistical inference to be *valid* and show that, in addition to the calibration property necessary for a meaningful interpretation of the inferential output, decision procedures with exact frequentist error rate control can easily be derived. For our second contribution, we demonstrate, in Section 3, that existing approaches, including frequentist and Bayesian, are not able to achieve our notion of valid statistical inference. This begs the question: *is valid statistical inference possible?* After a discussion of measuring degrees of belief in Section 4, our third contribution, in Section 5, is to prove that the *inferential model* (IM) framework (Martin and Liu 2013, 2015a,b,c) indeed provides valid statistical inference, according

---

<sup>1</sup><http://www.nature.com/nature/focus/reproducibility/index.html>

<sup>2</sup><http://amstat.tandfonline.com/doi/abs/10.1080/00031305.2016.1154108>

to our definition, and, therefore, provides a solution to what has been called “the most important unresolved problem in statistical inference” (Efron 2013b). The IM approach is based on the notion of predicting an unobservable auxiliary variable with a random set, and the resulting inferential output is a belief function, a departure from the familiar notion of probability and Bayesian or fiducial-type inference. Section 6 discusses some “principles” for statistical inference from this IM perspective, with an emphasis on marginalization, and some concluding remarks are given in Section 7.

## 2 What is statistical inference?

### 2.1 Setup, notation, and objectives

To start, we assume that there is a model  $P_{X|\theta}$ , depending on a parameter  $\theta \in \Theta$ , for the observable data  $X \in \mathbb{X}$ . No doubt, readers are familiar with statements like the previous one, but it contains (at least) two non-trivial concepts, namely, *data* and *model*. Here, as is common in discussions of inference, we assume that both the data and model are given, but we do have a few comments on each.

- *What is data?* Intuitively, data is what the statistician gets to see; in fact, the data is the only thing in a statistics problem that is really *known*. Here we use the adjective “observable” to distinguish the data  $X$  from other variables that are “unobservable;” see Sections 5 and 6.
- *What is a model?* Lehmann and Casella (1998, Ch. 1) define a model as a family of probability measures on  $\mathbb{X}$ , but this hides the difficulty and importance of specifying a good model. General discussion and guidelines on modeling are given, e.g., in Cox and Hinkley (1974), Box (1980), and McCullagh and Nelder (1983), and McCullagh (2002) gives a more technical discussion. Here it will suffice to consider the model  $P_{X|\theta}$ ,  $\theta \in \Theta$ , as describing data analyst’s view of how the observed data could be related to the scientific question of interest, derived from some knowledge of the data-generating process or from an exploratory data analysis.

The two questions mentioned above are fundamental in the sense that inference depends critically on the quality of the model and of the data collected. In light of the recent emergence of “big-data,” there may be a need for new papers and discussions about data and modeling, but we will leave this job to the experts.

Some statisticians, subjective Bayesians especially, might say that what we have called the model above is incomplete in the sense that it should also include an uncertainty assessment about the parameter  $\theta$  in the form of a prior; see, for example, de Finetti (1972), Savage (1972), Gelman et al. (2004), and Kadane (2011). If a real subjective prior is available, then the statistical inference problem is “simply an application of probability” (Kadane 2011, p. xxv), in particular, Bayes’s theorem (e.g., Efron 2013a) yields the conditional distribution of  $\theta$ , given  $X = x$ , which can be used for inference. We cannot criticize this subjective approach; in fact, if real subjective prior information is available, we recommend using it. However, there is an expanding collection of work (e.g., machine learning, etc) that takes the perspective that no real prior information is available. Even a large part of the literature claiming to be Bayesian has abandoned the

interpretation of the prior as a serious part of the model, opting for “default” prior that “works.” Our choice to omit a prior from the model is *not* for the (misleading) purpose of being “objective”—subjectivity is necessary—but, rather, for the purpose of exploring what can be done in cases where a fully satisfactory prior is not available, to see what improvements can be made over the status quo.

Given the model and observable data, the goal is to reach some conclusions about  $\theta$  based on the observed value  $X = x$ . Classically, the kinds of “conclusions” that are sought are point/interval estimates or hypothesis tests. To us, these are procedures and not really summaries of what information the data provides about the unknown parameter. The next subsection attempts to pin down what exactly is “inference.”

## 2.2 The definition

It is standard (e.g., Kadane 2011; Lindley 2014) to describe ones beliefs in the presence of uncertainty using probability. We propose to add a bit more flexibility and ask only that the uncertainty assessment be “probabilistic” in the sense that, for each assertion  $A \subseteq \Theta$  about the parameter  $\theta$ , the output assigns a numerical value that can be interpreted as ones (subjective) degree of belief in  $A$ , given data  $X$ , or, alternatively, the amount of evidence in data  $X$  supporting the truthfulness of  $A$ ; see Dempster (2014). This would include the familiar Bayesian approach, as well as others.

**Definition 1.** Given a model  $\{P_{X|\theta} : \theta \in \Theta\}$  for the observable data  $X$ , *statistical inference* is the construction of a function  $b_x : 2^\Theta \rightarrow [0, 1]$  such that, for each  $A \subseteq \Theta$ ,  $b_x(A)$  represents the data analyst’s degree of belief in the truthfulness of the claim “ $\theta \in A$ ” based on observation  $X = x$ .

This definition formalizes the idea that inference is the process of converting observations into degrees of belief about  $\theta$ . The output  $b_x$  in Definition 1 is a set-function, scaled to the interval  $[0, 1]$ , similar to a probability measure. However, we *do not* assume  $b_x$  satisfies all the properties of a probability measure. In particular, the definition does not require that we specify a prior, nor does it rule out the possibility that *both*  $b_x(A)$  and  $b_x(A^c)$  are small. This latter point will be discussed further in Section 4.

The inferential output  $b_x$  can be used in a very natural way. That is, large values of  $b_x(A)$  and  $b_x(A^c)$  suggest that data strongly supports the truthfulness and falsity, respectively, of the claim “ $\theta \in A$ .”<sup>3</sup> In this sense,  $b_x$  summarizes the data analyst’s uncertainty about  $\theta$  based on the the observed data  $X = x$ . Moreover, the data analyst can, if desired, produce decision procedures based on the inferential output, e.g., reject a null hypothesis “ $\theta \in A$ ” if and only if  $b_x(A^c)$  is sufficiently large; see Section 2.3. The data analyst’s beliefs are necessarily meaningful to him or her, but it remains to specify in what sense they are meaningful to others for scientific applications.

## 2.3 Validity condition

The reader is likely concerned about basing scientific conclusions on the data analyst’s “beliefs,” and rightfully so. Indeed, knowledge is gained when the evidence supporting a

---

<sup>3</sup>Intermediate values of  $b_x(A)$ , such as 0.4, are more difficult to interpret, but the same is true for probabilities, according to the Cournot principle (Shafer 2007; Shafer and Vovk 2006).

theory overwhelms the evidence to the contrary. This aggregation requires that experts can agree on whether observed data supports the truthfulness of an assertion, say, “ $\theta \in A$ ” and to what extent. In our context, this requires that the data analyst’s beliefs  $b_x$  be calibrated to a fixed and known scale. To this point, Reid and Cox (2015) write

*even if an empirical frequency-based view of probability is not used directly as a basis for inference, it is unacceptable if a procedure... of representing uncertain knowledge would, if used repeatedly, give systematically misleading conclusions.*

To avoid this “unacceptable” behavior, we propose the following calibration constraint on the inferential output.

**Definition 2.** Statistical inference based on  $b_x$  is *valid* if

$$\sup_{\theta \notin A} \mathbf{P}_{X|\theta} \{b_X(A) > 1 - \alpha\} \leq \alpha, \quad \forall \alpha \in [0, 1], \quad \forall A \subset \Theta. \quad (1)$$

That is, if the assertion  $A$  is false, then  $b_X(A)$ , as a function of  $X \sim \mathbf{P}_{X|\theta}$ , is stochastically no larger than  $\text{Unif}(0, 1)$ .

Intuitively, the validity condition (1) says that, if  $A$  is false, then there is a small probability that  $b_X(A)$  takes a large value or, in other words, data are unlikely to provide strong support for a false assertion, avoiding the “systematically misleading conclusions.” This is a desirable property that most existing methods lack; see Figure 2 in Section 5.2. Note that (1) covers *all* assertions, including singletons and their complements, not just convenient one-sided interval assertions, etc. Moreover, we are all familiar with how to interpret values on the  $\text{Unif}(0, 1)$  scale, and the condition (1) implies that this same scale can be used for interpreting the values of  $b_X(A)$ . See, also, Rubin (1984).

Validity is designed to make the data analyst’s inferential output meaningful to and interpretable by others. An interesting consequence is that procedures designed based on this inferential output will have guaranteed frequentist properties. Given the inferential output  $b_x$ , define the complementary function

$$p_x(A) = 1 - b_x(A^c), \quad A \subseteq \Theta,$$

which is, in general, different from  $b_x$ ; see Section 4.2. Since the validity condition holds for all  $A \subseteq \Theta$ , there is an equivalent formulation in terms of  $p_x$ :

$$\sup_{\theta \in A} \mathbf{P}_{X|\theta} \{p_X(A) \leq \alpha\} \leq \alpha, \quad \forall \alpha \in [0, 1], \quad \forall A \subseteq \Theta. \quad (2)$$

Besides complementing  $b_x$ ,  $p_x$  provides a simple and intuitive recipe to construct procedures with guaranteed frequentist error rate control.

**Theorem 1.** Let  $b_x$  be as in Definition 2 and fix  $\alpha \in (0, 1)$ .

- (a) Consider testing  $H_0 : \theta \in A$  versus  $H_1 : \theta \notin A$ . Then the test that rejects  $H_0$  if and only if  $p_x(A) \leq \alpha$  has frequentist Type I error probability upper bounded by  $\alpha$ .

(b) *The set*

$$\mathcal{P}_\alpha(x) = \{\vartheta \in \Theta : p_x(\{\vartheta\}) > \alpha\} \quad (3)$$

*has frequentist coverage probability lower-bounded by  $1 - \alpha$ , i.e.,  $\mathcal{P}_\alpha(x)$  is a nominal  $100(1 - \alpha)\%$  confidence region.*

The proof of the theorem follows immediately from the validity condition (1). Interestingly, there are no conditions on the sampling model and there is no need for asymptotic approximations. Therefore, in addition to calibrating the data analyst’s beliefs for interpretation by others, validity is also basically a necessary condition for all “good” frequentist methods. The question of whether valid statistical inference in the sense of the above definitions can be achieved will be addressed in what follows.

## 2.4 Remarks on foundations

According to our definition, statistical inference boils down to a conversion of the sampling model’s probabilities, which often have a frequency interpretation, to output having a belief/subjective interpretation. This is an important foundational observation: it makes clear the logical step one must take to go from a probability model for observable data to inference about the unknown parameter given the observed data. However, in scientific applications, it is important that the inferential output can be interpreted meaningfully by others, so “calibration seems essential” (Reid and Cox 2015, p. 295). As Evans (2015, p. xvi) writes, “subjectivity can never be avoided but its effects can be...controlled.” The validity condition in Definition 2 is designed specifically to strike a suitable balance between the unavoidable subjectivity in the inferential output and the calibration necessary for scientific applications.

Assuming one agrees with our claim that validity is essential, a natural follow-up question is how can it be achieved in practice. This, to us, is the fundamental question that the foundations of statistical inference should strive to answer. Indeed, those failures (e.g., irreproducibility, etc) quoted in Section 1 are all, at some level, consequences of a lack of validity so, to surely avoid these, data analysts need a set of guiding principles that, if followed, will guarantee that validity is achieved. This establishes the fundamental but under-appreciated connection between the foundations and practice of statistics. We must stress that the many “asymptotic validity” results, while technically interesting, apparently are not strong enough for all practical purposes—provable finite-sample validity is required. This is indeed a lofty goal, but our main foundational contribution here will be to demonstrate that provable validity is attainable.

To summarize our foundational position, we insist that statistical inference be provably valid, and it is from this perspective that we will evaluate existing methods. Among the different methods which are provably valid, of course, we recommend the one which is most efficient. Some comments about efficiency are made in Section 6 and 7, but our primary focus in this paper is on the necessary validity property.

## 2.5 Some historical perspective

The notion that inference requires an appropriate balance between subjectivity and “objectivity” can be found in the existing literature. In historical writings about Fisher, e.g.,



Aldrich (2000, 2008), Edwards (1997), and Zabell (1989, 1992), the emphasis is often on how his views about inverse and fiducial probability changed over the course of his career. Fisher’s shifts in perspective are problematic for those of us who hope to understand what he had in mind, but he should not be so severely criticized for this. To us, Fisher’s change in views—from Bayesian in his early years, to his original frequency-driven version of fiducial in the middle of his career, and to a conditional, more subjectivist version of fiducial near the end of his career—are particularly telling. Over time, it seems that Fisher found what are now the mainstream approaches in statistics to be unsatisfactory. Unfortunately, his last attempts to clarify and formalize the fiducial argument were unsuccessful, which led to fiducial being labeled Fisher’s “one great failure” (Zabell 1992, p. 369) and “biggest blunder” (Efron 1998, p. 105). Zabell (1992, p. 382) summarizes this beautifully:

*Fisher’s attempt to steer a path between the Scylla of unconditional behaviorist methods which disavow any attempt at “inference” and the Charybdis of subjectivism in science was founded on important concerns, and his personal failure to arrive at a satisfactory solution to the problem means only that the problem remains unsolved, not that it does not exist.*

Efron (2013b, p. 41) addresses this same point, highlighting the need for meaningful prior-free probabilistic inference:

*...perhaps the most important unresolved problem in statistical inference is the use of Bayes theorem in the absence of prior information.*

Similar comments can be found in Efron (1998), pointing to “something like fiducial inference” as a possible solution. There are now a number of variants of fiducial inference and, to us, what is missing is the satisfactory balance that Fisher was striving for, between the calibration properties of unconditional frequency-based methods and the subjective meaningfulness of conditional methods. Both considerations are important but, to date, no approach has addressed the two simultaneously. A main goal of this paper is to argue that the IM approach, described in Section 5, does strike this balance and, consequently, is a solution to Efron’s “most important unresolved problem.”

## 3 On standard modes of inference

### 3.1 Frequentist

The literature often refers to “frequentist methods” or a “frequentist approach.” To us, “frequentist” is not an approach to statistical inference but a method of evaluating a given method, e.g., one might ask what is the frequentist coverage probability of a Bayesian posterior credible interval. Our interpretation of “frequentist inference” is that which chooses a decision procedure based solely on its frequency properties. That is, no attempt is made to describe the data analyst’s degrees of belief. Based on this interpretation, “frequentist inference” is not inference in the sense of Definition 1:

*...what “frequentist” statisticians call “inference” is not inference in the natural language meaning of the word. The latter means to me direct situation-specific assessments of probabilistic uncertainties. (Dempster 2014, p. 267)*

The reader may find this high-level criticism difficult to swallow, but a more obvious and practical concern is that frequentism does not provide any guidance for selecting a particular rule or procedure. How, then, can frequentism yield a genuine framework for statistical inference? To us, the only real “frequentist approach” is one that selects a desirable feature, i.e., small mean-square error for estimators or small expected length of fixed-level confidence intervals, and then picks the best procedure according to this feature. Existence and uniqueness of this “best” procedure aside, there are examples where no procedures are good, making the search for “best” meaningless.

*Normal CV example.* Suppose that the observable data  $X$  is an iid sample of size  $n$  from a normal population,  $N(\mu, \sigma^2)$ , where both the mean  $\mu$  and the variance  $\sigma^2$  are unknown parameters. The goal is inference on  $\theta = \sigma/\mu$ , the coefficient of variation. Despite its apparent simplicity, this is a notoriously difficult problem for all existing approaches. In particular, it follows from Gleser and Hwang (1987, Theorem 1) that any confidence interval for  $\theta$  has positive coverage probability if and only if it is infinitely long with positive probability. This can lead to problems with interpretation, e.g., it is possible/likely to assign “95% confidence” to the interval  $(-\infty, \infty)$ , which is problematic. We must, therefore, conclude that frequentist inference, according to our interpretation, is questionable as a general approach/framework. There are other examples in this Gleser–Hwang class for which similar conclusions would be reached, including the famous Fieller–Creasy problem (Creasy 1954; Fieller 1954). Furthermore, Gleser and Hwang (1987, Sec. 5) point out that the use of asymptotically approximate confidence intervals in this class of examples does not resolve the problem, in fact, only exaggerates it more.

A few additional comments are in order. First, certain tools that fall under the umbrella of frequentist methods can be given a probabilistic inference interpretation. One example is the p-value, which can be understood as the “plausibility” of the null hypothesis, given data (Martin and Liu 2014b). We expect that other frequentist output, such as confidence intervals, can also be given a probabilistic inference interpretation, but the above example suggests there is lurking danger, so more work is needed. Second, repeated sampling properties are important for scientific inference, in particular, in terms of reproducibility. Our opinion is that procedures with good properties should be a consequence of meaningful and properly calibrated inferential output. So, we suggest that the work be carried out at the higher level of providing meaningful and efficient probabilistic inference, and then easily-derived decision procedures will automatically inherit the desirable frequentist properties as in Theorem 1.

## 3.2 Bayesian

In the absence of real subjective prior information, which is our context, standard Bayesian practice is to introduce a particular “default,” “objective,” or “non-informative” prior and apply the Bayes formula as usual. Key references on this approach include Kass and Wasserman (1996), Berger (2006), Berger et al. (2009), and Ghosh (2011). For smooth, one-dimensional, unconstrained problems, the Jeffreys prior is the agreed-upon choice, satisfying a variety of desirable objectivity properties (e.g., Ghosh et al. 2006, Ch. 5.1). For higher-dimensional problems, however, the choice of a good prior is less clear because different objectivity properties lead to different choices of prior, not to mention that “it leads to well-established difficulties with marginalization and with calibration” (Reid and Cox



2015); see, also, Section 6. Similar problems arise for constrained parameter problems, including one-dimensional ones.

Of the standard objectivity properties one might shoot for, the only one that could be appealing to us, given our emphasis on validity, is probability-matching, i.e., selecting the prior so that the posterior credible intervals have (approximately) the nominal frequentist coverage probability. But, as the following example demonstrates, there are cases where probability-matching is not possible, raising the question of whether the objective Bayesian approach can be considered as a general framework for scientific inference.

*Normal CV example* (cont). Berger et al. (1999, Example 7) discuss several likelihood and Bayesian approaches to inference on the normal coefficient of variation,  $\theta = \sigma/\mu$ . By introducing a marginal reference prior for  $\theta$ , Equation (38) in their paper, they obtain a proper posterior distribution in their subsequent Equation (39). However, this proper posterior will yield credible intervals with finite length and, according to the Gleser–Hwang theorem, must have poor frequentist coverage probability. The same conclusion can be reached for any prior leading to a proper posterior and, therefore, probability-matching is not possible in this example. This undesirable mis-calibration carries over to other features of the posterior besides credible regions. Indeed, Figure 2 in Section 5.2 gives an example where the posterior probability assigned to a false assertion is near 1 in almost every simulation run—a “systematically misleading conclusion.”

Ultimately, our main concerns about the default-prior Bayes approach revolve around the interpretation of the posterior probabilities. If the prior has a frequency or belief probability interpretation, then the posterior inherits that. However, a default prior has neither interpretation, so then it is not clear in what sense the posterior probabilities should be interpreted.

*[Bayes’s formula] does not create real probabilities from hypothetical probabilities... (Fraser 2014, p. 249)*

A related issue is the scale of the posterior probabilities. Mathematically, the posterior is absolutely continuous with respect to the prior, i.e., any event with zero prior probability also has zero posterior probability. So, data cannot provide support for any assertion that corresponds to an event with prior probability zero; this leads to practical difficulties in Bayesian testing point null hypotheses. More generally, those events with small prior probability will also have small posterior probability, unless the information in data strongly contradicts that in the prior. From this point of view, it is clear that the posterior is only as meaningful as the prior, at least for moderately informative data. Therefore, we agree with the assessment of Bayesian approach as just “quick and dirty confidence” (Fraser 2011), i.e., it is a conceptually straightforward and general strategy for constructing procedures with good (asymptotic) frequency properties.

### 3.3 Fiducial and related ideas

The methods to be considered here fall under the general umbrella of “distributional inference,” where the inferential output is a data-dependent probability distribution on the parameter space. Besides the Bayesian approach discussed above, this also includes fiducial inference (Barnard 1995; Fisher 1973; Zabell 1992), structural inference (Fraser

1968), generalized inference (Chiang 2001; Weerahandi 1993), generalized fiducial inference (E et al. 2008; Hannig 2009, 2013; Hannig et al. 2006; Hannig and Lee 2009; Lai et al. 2015; Wang et al. 2012), confidence distributions (Schweder and Hjort 2002; Xie and Singh 2013), and Bayesian inference with data-dependent priors (Fraser et al. 2010; Martin and Walker 2014). As Efron (1998, p. 107) famously said,

*Maybe Fisher’s biggest blunder will become a hit in the 21st century!*

True to Efron’s prediction, these fiducial-like methods have been applied successfully in a variety of interesting practical problems. Indeed, confidence distributions have proved to be useful tools in meta-analysis applications (Liu et al. 2015; Xie et al. 2011); data-dependent priors in a Bayesian analysis have shown to be beneficial in high-dimensional problems or when higher-order accuracy is required; and an interesting general observation is that the generalized fiducial methods often perform “too well,” see, e.g., E et al. (2008) and Hannig et al. (2015), a phenomenon that has yet to be explained.

We should differentiate this sort of distributional inference from our notion of probabilistic inference in Definition 1. The former makes no claim that the “posterior probability” of an assertion  $A \subseteq \Theta$  is a meaningful summary of the evidence in data supporting the truthfulness of  $A$ . In fact, these methods are typically used only to construct interval estimates for  $\theta$  by returning, say, the middle 95% of the posterior distribution, and relying on asymptotic theory to justify the claimed “95% confidence.” However, as the next example shows, constructing a genuine distribution estimate is not always possible, casting doubt on the potential of this class of methods providing a fully satisfactory framework for statistical inference.

*Normal CV example (cont).* Dempster (1963) demonstrated that there is a non-uniqueness issue in constructing a joint fiducial posterior distribution for  $(\mu, \sigma^2)$  in the normal problem, which would obviously persist in the marginal fiducial distribution of  $\theta = \sigma/\mu$ . Hannig (2009) gives a generalized fiducial distribution for  $(\mu, \sigma^2)$ , which agrees with the standard default-prior Bayes solution, and a corresponding fiducial distribution for  $\theta$  obtains by integration. However, based on our arguments in Section 3.2, this marginal fiducial distribution for  $\theta$  cannot have a meaningful interpretation. Hannig, in a personal communication, has indicated that integrating the joint generalized fiducial distribution for  $(\mu, \sigma^2)$  to get marginal distribution for  $\theta$  is not appropriate in this case, and that he would carry out his construction differently.

A confidence distribution for  $\theta$  would obtain in this case by setting the quantiles equal to the endpoints of a set of suitable upper confidence limits. By the Gleser–Hwang theorem, with positive probability, these upper limits will all be  $\infty$ , making the corresponding confidence distribution degenerate there. So, the standard approach of stacking one-sided confidence intervals to get a distribution for  $\theta$  will not be successful in this example. An alternative construction of a “confidence curve” is given by Schweder and Hjort (2016, Sec. 4.6) for the related Fieller–Creasy problem.

We agree that a distribution estimate provides more information than a point estimate or a confidence interval. However, the proposed use of these distributions is largely to read off interval estimates or p-values, and these typically only have a meaningful interpretation asymptotically. Moreover, fiducial and confidence distributions cannot be manipulated like genuine distributions, i.e., marginalization via integration generally will

fail. So it is not clear to us what is the benefit of having a “distribution;” see, also, Robert (2013). Therefore, we classify these distributional inference methods as (useful) tools for designing frequentist procedures and, consequently, these are subject to the remarks in Fraser (2011) concerning Bayes and confidence.

## 4 Measuring degrees of belief

### 4.1 Example: describing ignorance

As an illustrative example of degrees of belief, we consider the extreme case of *ignorance*, or lack of knowledge. Despite being extreme, ignorance is apparently a practically relevant case since the developments of default priors are motivated by examples where the data analyst is ignorant about the parameter, i.e., no genuine prior information is available. So, understanding how to describe ignorance mathematically ought to be insightful for the construction of meaningful prior-free probabilistic inference.

What does ignorance mean? A standard assumption is that the universe  $\Theta$ , the parameter space, is given—so we are not *totally ignorant*—and, in order to make the problem interesting, we assume that  $\Theta$  contains at least two points. Subject to these conditions, ignorance means that there is no evidence available that supports the truthfulness or falsity of any assertion  $A \subset \Theta$ , where  $A \neq \emptyset, \Theta$ . Then the natural way to encode this mathematically is via the function  $b : 2^\Theta \rightarrow [0, 1]$  such that  $b(A) = 0$  for all proper subsets  $A$  of  $\Theta$ . Shafer (1976) and others call this a *vacuous* belief.

For illustration, suppose that we are interested in the weekday  $\theta$  (Sunday–Saturday) on which you, the reader, were born. All seven days are plausible but we have no evidence to support the truthfulness of a claim “ $\theta \in A$ ” for any proper subset  $A$  of the possible weekdays. Therefore, we may choose to specify our beliefs based on the ignorance model described above. Compare this to the arguably more standard approach where a *probability* of  $1/7$  is assigned to each of the possible weekdays. Such a summary is based on a judgement that the uncertainty about  $\theta$  is equivalent to the uncertainty about the outcome of an experiment that rolls a fair seven-sided die, and the latter is based on a judgement of symmetry, i.e., knowledge. Therefore, it is clear that the equi-probable model does not describe the state of ignorance.

Bayes and Laplace, more than 200 years ago, employed a uniform prior for  $\theta$  in the binomial model based on a “principle of indifference,” i.e., in a partition of  $[0, 1]$  into intervals of equal length, all are equally likely to contain  $\theta$ . However, the above reasoning still applies and one can readily conclude that the “indifferent” uniform prior cannot represent the state of ignorance. The point is that even indifference is a type of knowledge. More generally, an easy by-contradiction argument shows that no standard prior, proper or improper, can encode ignorance. Therefore, no standard Bayesian prior can accommodate the data analyst’s ignorance about  $\theta$ , i.e., the prior must be introducing some knowledge. This begs the question: can a prior really be “non-informative?”

Professor Jim Berger asked one of us (CL) an important question following a workshop session in 2015:<sup>4</sup> “isn’t objective Bayes prior-free?” It is true that, in carrying out an objective Bayes analysis, elicitation of subjective prior information is not required,

---

<sup>4</sup><http://www.bayes.ecnu.edu.cn/BFF2015>

but this alone does not make the analysis prior-free. If the prior does not reflect the data analyst’s state of ignorance, then the corresponding posterior is artificial—recall the quotes from Fraser in Section 3.2. We have argued that no prior can encode the state of ignorance, so the use of default priors introduces some artificiality and, therefore, the corresponding analysis cannot be considered prior-free.

To be clear, our intention here is not to advocate for the state of ignorance. The goal was simply to consider how ignorance can be described mathematically and to consider if a Bayesian prior is able to accomplish this. The discussion here reveals that, if no prior information is available—an assumption that the majority of statisticians make—then, to describe the state of ignorance and to avoid introducing some artificiality, the data analyst must depart from the familiar framework of probability theory. This, along with the discussion in Section 4.2 below, provides motivation for our use of belief functions.

## 4.2 Additivity and evidence

Probability is predictive in nature, designed to describe uncertainties about a yet-to-be-observed outcome. In the statistical inference problem, however,  $\theta$  is a fixed but unknown quantity, never to be observed, so there is no reason to require the summaries of our uncertainty to satisfy the properties of a probability measure.

Let  $\Pi_X$  be a distribution estimator of  $\theta$  given data  $X$ , i.e., a Bayesian posterior distribution, a (generalized) fiducial distribution, or a confidence distribution. For a fixed assertion  $A \subseteq \Theta$ , the posterior probability  $\Pi_X(A)$  is to be interpreted as a measure of the evidence in  $X$  supporting the truthfulness of the claim “ $\theta \in A$ .” Additivity of  $\Pi_X$  implies that the measure of evidence in  $X$  supporting the falsity of that same claim is  $\Pi_X(A^c) = 1 - \Pi_X(A)$ . The problem, a consequence of additivity, is that there is no middle ground. In other words, evidence in  $X$  supporting the truthfulness of “ $\theta \in A$ ” is necessarily evidence supporting the falsity of its negation. If  $\theta$  were a yet-to-be-observed outcome from some experiment, then the dichotomy of  $\theta$  *is in exactly one of  $A$  and  $A^c$*  is perfectly fine. When dealing with evidence, however, no such dichotomy exists. For example, consider a criminal court trial, and let  $A$  be the assertion that the defendant is not guilty. Witness testimony that corroborates the defendant’s alibi is evidence that supports the truthfulness of  $A$  but it does not provide direct evidence to support the falsity of  $A^c$ . So, there ought to be a middle-ground “don’t know” category (Dempster 2008) to which some evidence can be assigned, leading to a super-additive summary of evidence. Dempster (2014, Sec. 24.3) points out that “don’t know” probabilities accommodate the certain inadequacies of the posited model ignored by standard approaches.

To further illustrate the need for a “don’t know” category, let us take a simple example, i.e.,  $X \sim N(\theta, 1)$ . With a flat prior for  $\theta$ , the posterior distribution for  $\theta$ , given  $X$ , is  $N(X, 1)$ . Let  $A = (-\infty, \theta_0]$ , for fixed  $\theta_0$ , be the assertion of interest. Consider the extreme case where  $X$  is close to  $\theta_0$ , say,  $X = \theta_0 - \varepsilon$ , for a small constant  $\varepsilon > 0$ . In such a case,  $\Pi_X(A)$  and  $\Pi_X(A^c)$  are both roughly 0.5. Certainly,  $X$  sitting on the boundary cannot distinguish between  $A$  and  $A^c$ , and the equally probable conclusion reached by the Bayesian posterior distribution is consistent with this. Our concern, however, is about the magnitude of the value 0.5. As a measure of support, the value 0.5 suggests that there is non-negligible evidence in data supporting the truthfulness of both  $A$  and  $A^c$ , which is counterintuitive—a single observation at the boundary should not provide

strong support for either assertion. In this sense, it is more reasonable to assign small values to both  $A$  and  $A^c$ , communicating the point that the single data point provides minimal support to both. Again, this suggests that the evidence measure should be super-additive. If the data were more informative, e.g., if the sample mean of 1000 observations equals  $\theta_0 - \varepsilon$ , then the posterior summary would assign values 0 and 1, roughly, to  $A$  and  $A^c$ , respectively, and we agree that these would be meaningful and appropriately scaled summaries of the evidence in the data. So, that third “don’t know” category should disappear as data become more informative, leaving a probability measure as the asymptotic summary of evidence. This helps to further explain why those distributional inference methods (Bayes, fiducial, etc) can provide valid summaries only asymptotically: they are large-sample approximations of a meaningful belief function.

### 4.3 Belief functions

In the previous sections, we have hinted that belief functions are the appropriate tools for describing degrees of belief about the parameter after data has been observed. One example of a belief function has already been presented—the so-called “vacuous” belief function in Section 4.1—but here we provide some more detailed background and make an important connection to random sets. For further details and applications, the reader is referred to Shafer (1976) or Yager and Liu (2008).

The standard presentation of belief functions considers the case of a finite space  $\Theta$ . Start with a probability mass function  $m$  supported on  $2^\Theta$  with the property that  $m(\emptyset) = 0$ . For  $A \subseteq \Theta$ , the quantity  $m(A)$  encodes the degree of belief in the truthfulness of  $A$  but in no proper subset of  $A$ . Then the *belief function* can be defined as

$$b(A) = \sum_{B: B \subseteq A} m(B), \quad A \subseteq \Theta. \quad (4)$$

That is, the belief in the truthfulness of the assertion  $A$  is defined as the totality of the degrees of belief in  $A$  and all its proper subsets. Despite only covering the relatively simple finite case, this formula is insightful. In particular, the claimed super-additivity of belief functions can be seen immediately from the formula above: in general, there are some sets  $B \subseteq \Theta$  that are subsets of neither  $A$  nor  $A^c$ , and  $b(A) + b(A^c)$  does not count the mass assigned by  $m$  to these subsets. A related quantity, that will appear again in our discussion in Section 5.1, is the *plausibility function*, i.e.,

$$p(A) = 1 - b(A^c), \quad A \subseteq \Theta,$$

and  $p(A)$  represents the totality of the degrees of belief that do not contradict  $A$ . It is clear from the belief function’s super-additivity property that

$$b(A) \leq p(A), \quad A \subseteq \Theta,$$

with equality for all  $A$  if and only if  $b$  is a probability measure. This makes intuitive sense as well: an assertion is always at least as plausible as it is believable. Moreover, the difference  $p(A) - b(A)$  is the “don’t know” component (Dempster 2008), discussed in Section 4.2, responsible for providing an honest assessment of uncertainty, even in the extreme case of total ignorance.



We are not the first to consider the use of belief functions for statistical inference. However, to our knowledge, we are the first to recognize that the belief function’s properties are necessary in order for the inferential output to satisfy the required validity property (1). Indeed, with belief functions, validity at *all* assertions possible, even at the scientifically relevant singleton assertions and their complements.

A belief function can be defined on an arbitrary set  $\Theta$ . Indeed, as in Shafer (1979) or Wasserman (1990), a function  $b : 2^\Theta \rightarrow [0, 1]$  is a belief function if  $b(\emptyset) = 0$ ,  $b(\Theta) = 1$ , and it is monotone of order  $\infty$  in the sense of Choquet (1954); see, also, Nguyen (1978). This technical definition is not particularly insightful, however. Some of the first applications of belief functions were based on a sort of push-forward mapping of a probability measure via a set-valued mapping or, in other words, a *random set* (Molchanov 2005; Nguyen 2006). In fact,  $b(A)$  in (4) can easily be seen to be the probability that a random subset of  $\Theta$ , with mass function  $m$ , is a subset of  $A$ . Shafer (1979) demonstrated that, in a certain sense, all belief functions can be described via the distribution of a random set, simplifying their definition and construction. Random sets are an essential element the proposed IM framework, and some details are given in Section 5.1 below.

## 5 Valid prior-free probabilistic inference

### 5.1 Inferential models

The previous sections have challenged the status quo by proposing a serious definition of statistical inference and arguing that the existing approaches fall short of this. An obvious question is if there is an approach that meets our specified criteria. The recently proposed *inferential model* (IM) approach provides an affirmative answer to this question. Here we give a high-level introduction to the IM approach, as it relates to our notion of probabilistic inference; readers interested in the details of the IM construction and its properties are referred to Martin and Liu (2015b) and the references therein.

The starting point is the notion of valid probabilistic inference from Section 2. To satisfy the validity condition (1), it is clear that the inferential output  $b_X$  must depend on the sampling model somehow. One way to accommodate this is to select a particular representation of the data-generating process, i.e.,

$$X = a(\theta, U), \quad U \sim P_U, \tag{5}$$

where  $U \in \mathbb{U}$  is called an *auxiliary variable* with distribution  $P_U$  fully known, and  $a : \Theta \times \mathbb{U} \rightarrow \mathbb{X}$  is a known mapping. The representation in (5) is called an *association*. An association always exists since any sampling model that can be simulated on a computer must have a representation (5), but it is not unique. The association need not be based on knowledge of the actual data-generating process, it can be viewed simply as a representation of the data analyst’s uncertainty; that is, no claim that “ $U$ ” actually exists is required. The technical role played by the representation (5) is in identifying a fixed probability space—namely,  $(\mathbb{U}, P_U)$ —on which to carry out the probability calculations that will define  $b_X$ ; see (7).

Readers familiar with Fisher’s fiducial argument (e.g., Fisher 1973; Zabell 1992) will recognize the auxiliary variable  $U$  as a “pivotal quantity;” similar considerations can be



found in the structural inference of Fraser (1968). Roughly speaking, these approaches solve the equation (5) for  $\theta$  and use the distribution  $U \sim P_U$  and the observed value of  $X$  to induce a distribution for  $\theta$ . However, the resulting fiducial probabilities generally are not valid in the sense of (1); for details justifying this claim, see Martin and Liu (2015b, Chapter 2.3.2). The basic problem with the fiducial argument is that  $X$  and  $U$  in (5) are too tightly linked together, so we consider an alternative perspective.

According to the prescription in (5), there is a quantity  $U$  that, together with  $\theta$ , determines the value of  $X$ . If both  $X$  and  $U$  were observable, then we could simply solve for  $\theta$  and there would be no uncertainty. However, since  $U$  is not observable, the best we can do is to accurately *predict*<sup>5</sup> the unobserved value using some information about the distribution  $P_U$  that produced it. The point is that  $X$  is tied to the unobserved value of  $U$ , but *it is not tied to our prediction of that unobserved value*. This is precisely where our IM approach breaks with the fiducial argument; see, also, Remarks 1 and 3.

To accurately predict the unobserved value of  $U$ , we recommend the use of a random set, say  $\mathcal{S}$ , with distribution  $P_{\mathcal{S}}$ , with the property that it contains a large  $P_U$ -proportion of  $U$  values with high probability. In particular, if we set  $f_{\mathcal{S}}(u) = P_{\mathcal{S}}(\mathcal{S} \ni u)$ , then we require that

$$f_{\mathcal{S}}(U) \text{ is stochastically no smaller than } \text{Unif}(0, 1). \quad (6)$$

Intuitively, this condition suggests that  $\mathcal{S}$  can accurately predict a value  $U$  from  $P_U$ . Martin and Liu (2013) demonstrate that this is a relatively mild condition and provide some general strategies for constructing such a set  $\mathcal{S}$ ; see, e.g., (8).

*Remark 1.* One can alternatively understand the role played by the random set  $\mathcal{S}$  is to facilitate integration over the  $U$ -space *before* the observation  $X$  is conditioned on. The original fiducial argument proposes carrying out integration *after* conditioning on data  $X$ . This requires a questionable “continue to regard” assumption (Dempster 1963) that is responsible for many of the problems associated with fiducial inference. This provides a more mathematical explanation of how the IM approach differs from fiducial compared to the intuition-based arguments presented above. See, also, Remark 3.

Now, if  $\mathcal{S}$  provides a believable guess of where the unobserved value of  $U$  resides, in the sense of (6), then an equally believable guess of where the unknown  $\theta$  resides, given  $X$ , is  $\Theta_X(\mathcal{S}) = \bigcup_{u \in \mathcal{S}} \{\theta : X = a(\theta, u)\}$ , another random set, with distribution determined by  $P_{\mathcal{S}}$ . It is, therefore, natural to say that data  $X$  supports the claim “ $\theta \in A$ ” if  $\Theta_X(\mathcal{S})$  is a subset of  $A$ . The IM’s inferential output is

$$b_X(A) = P_{\mathcal{S}}\{\Theta_X(\mathcal{S}) \subseteq A\}, \quad (7)$$

which is the random set equivalent to the distribution function of a random variable. More precisely,  $b_X$  is a belief function, as discussed in Section 4.3, so it inherits those properties. Moreover, this IM construction, driven by the belief probability  $P_{\mathcal{S}}$  that satisfies (6), leads to valid probabilistic inference in the sense of Section 2.

**Theorem 2** (Martin and Liu 2013). *Suppose that  $\mathcal{S}$  satisfies (6) and that  $\Theta_x(\mathcal{S})$  is non-empty with  $P_{\mathcal{S}}$ -probability 1 for each  $x$ . Then the belief function  $b_X$  defined in (7) satisfies the validity condition (1).*

---

<sup>5</sup>Perhaps *guess* is a better word to describe this operation than *predict* but the former has a particularly non-scientific connotation, which we have opted to avoid.

The non-emptiness assumption about  $\Theta_x(\mathcal{S})$  in the theorem often holds automatically for natural choices of  $\mathcal{S}$ , but can fail in problems that involve non-trivial parameter constraints. In the case  $\mathbb{P}_{\mathcal{S}}\{\Theta_x(\mathcal{S}) \neq \emptyset\} < 1$ , Ermini Leaf and Liu (2012) propose a remedy that uses an *elastic* version of  $\mathcal{S}$ , one that will stretch just enough that  $\Theta_x(\mathcal{S})$  is non-empty, and they prove a corresponding validity theorem.

Recall that the validity condition was designed to properly calibrate the data analyst’s beliefs for scientific application. In Section 3 we argued that existing approaches to inference cannot achieve this calibration property, but Theorem 2 shows that this is easy to accomplish in the IM framework. An important consequence of the validity condition, presented in Theorem 1, is that it is straightforward to construct decision procedures with guaranteed frequentist error rate control. For example, the IM’s plausibility function,  $p_X(A) = 1 - b_X(A^c)$ , defines a plausibility region, as in (3), with the nominal frequentist coverage probability. And these desirable properties do not require any assumptions on the model or any asymptotic approximations. The take-away message is that the IM approach provides valid prior-free probabilistic inference and, therefore, solves Efron’s “most important unresolved problem.”

## 5.2 Normal coefficient of variation example, revisited

Reconsider the normal coefficient of variation example from Section 3. According to Martin and Liu (2015a, Sec. 4), we may first reduce to the minimal sufficient statistic  $(\bar{X}, S^2)$ , the sample mean and sample variance, respectively. Based on this pair, the (conditional) association can be written as

$$\begin{aligned} \bar{X} &= \mu + \sigma n^{-1/2} U_1, & U_1 &\sim \mathbf{N}(0, 1), \\ S &= \sigma U_2, & (n-1)U_2^2 &\sim \text{ChiSq}(n-1), \end{aligned}$$

where  $U_1$  and  $U_2$  are independent. The parameter of interest,  $\theta = \sigma/\mu$ , is a scalar but the auxiliary variable  $(U_1, U_2)$  is two-dimensional, so we would like to further reduce the dimension of the latter. Following the dimension-reduction arguments in Martin and Liu (2015c), we get a marginal association for  $\theta$ :

$$\frac{n^{1/2}\bar{X}}{S} = F_{n,1/\theta}^{-1}(U), \quad U \sim \text{Unif}(0, 1),$$

where  $F_{n,\psi}$  is the non-central Student-t distribution function, with  $n-1$  degrees of freedom and non-centrality parameter  $n^{1/2}\psi$ . The above expression corresponds to (5) in the general setup above. For simplicity, we follow Martin and Liu (2013) and consider a “default” random set for predicting the unobserved value of  $U$ , i.e.,

$$\mathcal{S} = \{u \in (0, 1) : |u - 0.5| \leq |U - 0.5|\}, \quad U \sim \text{Unif}(0, 1). \quad (8)$$

This predictive random set satisfies the sufficient condition (6) for the corresponding IM to be valid (Martin and Liu 2013). This yields the data-dependent random set

$$\Theta_X(\mathcal{S}) \equiv \bigcup_{u \in \mathcal{S}} \{\theta : F_{n,1/\theta}(t_X) = u\} = \{\theta : |F_{n,1/\theta}(t_X) - 0.5| \leq |U - 0.5|\},$$

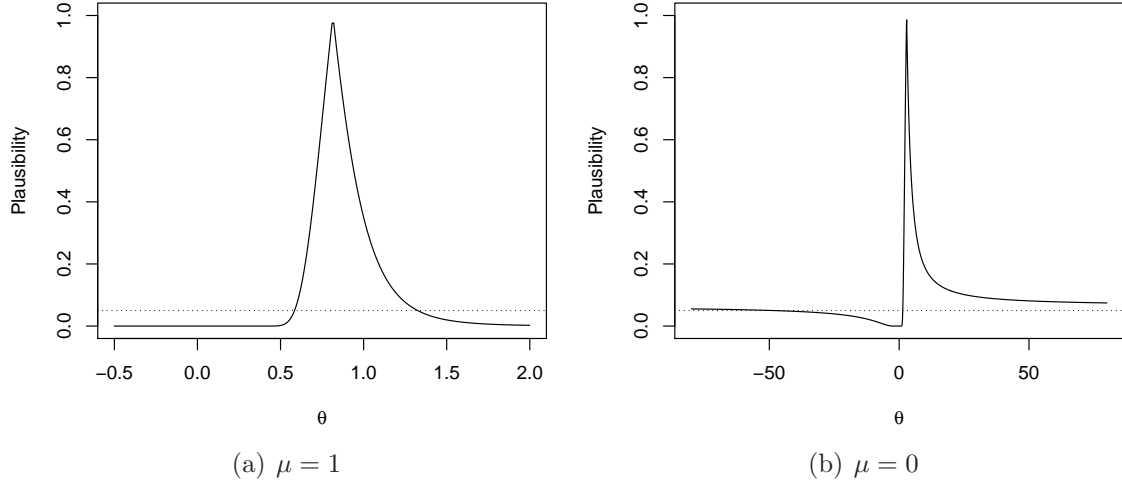


Figure 1: Plots of the plausibility function for  $\theta = \sigma/\mu$  based on samples of size  $n = 30$  from  $\mathbf{N}(\mu, \sigma^2)$  for  $\sigma = 1$  and two values of  $\mu$ .

where  $U \sim \text{Unif}(0, 1)$  and  $t_X = n^{1/2}\bar{X}/S$ . Since the IM is valid, one can produce valid probabilistic inference for any assertion about  $\theta$ . In particular, for singleton assertions, the belief function is zero but the plausibility function is

$$p_X(\{\theta\}) = 1 - |2F_{n,1/\theta}(t_X) - 1|.$$

We may construct an interval estimate for  $\theta$  using this plausibility function. Specifically, for fixed  $\alpha \in (0, 1)$ , the  $100(1 - \alpha)\%$  plausibility interval for  $\theta$  is  $\{\theta : p_X(\{\theta\}) > \alpha\}$  which, in this case, simplifies to

$$\{\theta : p_X(\{\theta\}) > \alpha\} = \{\theta : \alpha/2 < F_{n,1/\theta}(t_X) < 1 - \alpha/2\}.$$

This plausibility interval clearly has frequentist coverage probability  $1 - \alpha$ , which is a consequence of our general IM validity results.

It will be helpful to compare the output of our IM approach in this example to that from existing methods discussed in Section 3. The Gleser–Hwang theorem applies to our plausibility intervals as well, so these will be unbounded with positive probability; see Figure 1, which shows two plausibility functions, one that vanishes in the tails, yielding bounded plausibility intervals, and one that does not vanish in the tails, yielding unbounded plausibility regions. To us, however, this is not a concern: given the inherent variability of the sample mean around  $\mu$ , if  $\mu$  is close to zero, then arbitrarily large values of  $\theta$  cannot be ruled out. Our main point is that the IM output contains more information than just a plausibility interval/region. Indeed, the IM output is a valid belief/plausibility function pair that can be evaluated at any assertion  $A$  of interest, allowing the user to assess the evidence in the observed data supporting the truthfulness of  $A$  or  $A^c$ . The default-prior Bayesian posterior can be evaluated at these assertions but, as made clear in Section 3.2, the output is not meaningful in either a frequency or subjective sense.

To further illustrate the latter point about the difficulties in interpreting the Bayesian posterior probabilities, consider a simple example. Let  $X = (X_1, \dots, X_n)$  be an iid sample

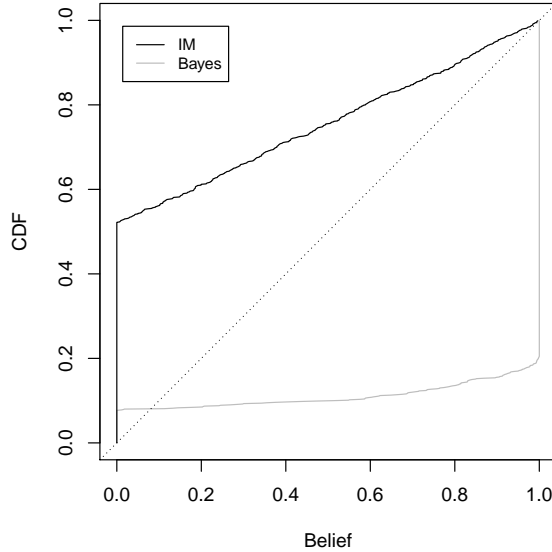


Figure 2: Quantile plot of the IM belief function  $b_X(A)$  and the default-prior Bayes posterior probability  $\Pi_X(A)$ , as functions of the iid sample  $X = (X_1, \dots, X_{10})$  from  $N(0.1, 1)$ , where  $A = (-\infty, 9]$ , which does not contain the true  $\theta = 10$ . Clearly, the Bayes posterior probability tends to be too large.

of size  $n = 10$  from  $N(\mu, \sigma^2)$ , with  $\mu = 0.1$  and  $\sigma = 1$ ; then the true value of  $\theta$  is 10. Now suppose the assertion of interest is  $A = (-\infty, 9]$ , which happens to be false. We can compare the IM belief function at  $A$ , based on the construction above, and the default-prior Bayes posterior probability of  $A$ , based on Equation (39) in Berger et al. (1999), as functions of  $X$ . Figure 2 shows the distribution functions of the belief function  $b_X(A)$  and the posterior probability  $\Pi_X(A)$  based on 1000 data sets. That the former is above the diagonal line is a demonstration of the validity property (1). The latter,  $\Pi_X(A)$ , is not valid because tends to be too large, even though  $\theta \notin A$ . Since large  $\Pi_X(A)$  suggests  $A$  is true, the default-prior Bayes inference, at least in this example, leads to “systematically misleading conclusions” and is, therefore, questionable.

*Remark 2.* In the above example, efficiency can be gained, i.e., the belief function line in Figure 2 can be made closer to the diagonal, by using a random set that is “optimal” for the assertion  $A$  (Martin and Liu 2013, Theorem 4). But improving the efficiency for inference at  $A$  in this way results in a loss of efficiency for inference at  $A^c$ . The “default” two-sided random set (8) effectively balances the efficiency across  $A$  and  $A^c$ .

## 6 “Principles”—an IM perspective

Statistical inference is an important and difficult problem, so the need for a set of guiding principles is clear, but its ill-posed nature suggests that a simple yet satisfactory set of mathematical axioms is unlikely. Case in point is Birnbaum’s theorem on the likelihood principle (Birnbaum 1962), a strong mathematical result in support of a Bayesian approach (e.g., Berger and Wolpert 1984), which has recently been refuted; see Evans

(2013), Mayo (2014), and the corresponding discussion. Based on the presentation in Reid and Cox (2015), perhaps the most recent discussion on “statistical principles,” the modern emphasis is on sound intuition as opposed to mathematical axioms, etc. On this we agree, and here we present some insights coming from our IM perspective.

To be clear, the principles to be described in this section are to provide sound intuition and not intended to be precise mathematical statements. Evans (2015, p. 1) writes that “attempts at mathematical generality often mislead as to what is appropriate statistical reasoning.” Perhaps the existing mathematics is not sufficient to describe a set of statistical principles, but it remains an interesting question if our intuition can be encoded in a set of simple yet precise mathematical axioms.

As we discussed in Section 5.1, the jumping off point of the IM approach is the specification of an association (5) that links an unobservable auxiliary variable  $U$  to the observable data  $X$  and parameter  $\theta$ . *A priori*, we can think of  $U$  as being responsible for the inherent variability in  $X$ , given  $\theta$ ; this is the view of (5) from the perspective of simulation. However, *a posteriori*, when data  $X$  is observed, our understanding of  $U$  should change. Specifically,  $U$  changes from a random variable to what we call a *predictable quantity*, that is, a fixed but unobserved value of a random variable whose distribution is completely known.

*Remark 3.* As mentioned in Section 5.1, the basic fiducial argument breaks down because it fails to recognize that the nature of  $U$  changes significantly after the data are conditioned on. Specifically,  $U$  is unpredictable after data are conditioned on. Indeed, the conditional distribution of  $U$ , given  $X$ , is degenerate on a solution to the equation in (5), which is not *known* because it depends on the true  $\theta$ .

Given a predictable quantity, it is possible, then, to define a random set  $\mathcal{S}$  for predicting that predictable quantity such that the condition (6) is satisfied. This leads to the following two principles, introduced in Martin and Liu (2014a):

**VALIDITY PRINCIPLE.** *Predict the predictable quantity with a random set satisfying (6).*

**EFFICIENCY PRINCIPLE.** *Subject to the constraint (6), the prediction of the predictable quantity should be done as efficiently as possible, in some sense.*

The validity principle is technically clear: according to the key results in Martin and Liu (2013), it basically reiterates our insistence in Section 2 that the inferential output be valid. The efficiency principle, on the other hand, is (purposely) vague, but it provides some sound intuition and new insights. In particular, the dimension of the predictable quantity  $U$  should be reduced as much as possible prior to introducing the random set. In this sense, the results in Martin and Liu (2015a,c) can be considered as recipes for putting the efficiency principle into action. The former paper shows that, in cases where the sample size exceeds the dimension of the parameter, certain features of the predictable quantity are observed, suggesting a dimension reduction strategy that includes Fisher’s notions of sufficiency and conditioning on relevant subsets as special cases. Moreover, if genuine prior information is available, then it is possible to condition on all of the data, and the paper shows that the IM approach reproduces the Bayesian solution. The latter paper considers marginal inference and provides a characterization of the class of problems in which certain features of the predictable quantity  $U$  can be linked to nuisance parameters, suggesting a dimension reduction by simply ignoring these features. This

characterization of “regular” marginal inference problems has proved to be useful in an unrelated context; see Theorem 2.3 in Hannig et al. (2015).

Marginalization is an important problem, one that received considerable attention at a recent “statistical foundations” workshop at Rutgers University;<sup>6</sup> in fact, the presentations given by both D. R. Cox and Don Fraser focused primarily on marginalization issues. Given its importance, it may help to elaborate a bit more on the IM approach to marginalization. To give the discussion context, consider the length problem of Stein (1959), i.e., that of estimating the squared length  $\theta = \|\mu\|^2$  of the normal mean vector  $\mu \in \mathbb{R}^n$  based on a single sample  $X \sim \mathbf{N}_n(\mu, I)$ , where  $I$  is the  $n \times n$  identity matrix. The paradox appears because the “obvious” approaches to this problem fail miserably. For example, the maximum likelihood estimator of  $\theta$  has substantial bias, and the Bayesian posterior for  $\theta$  obtained by taking a flat prior for  $\mu$ , applying Bayes formula, and then marginalizing is severely mis-calibrated; this mis-calibration problem exists also for the fiducial distribution for  $\theta$  obtained by directly marginalizing the fiducial distribution for  $\mu$ . The common remedy for both the Bayesian and fiducial mis-calibration is to bring the marginal nature of the problem into the picture *before* deriving the posterior. For example, if  $\theta$  is of primary interest, the Bayesian can construct a posterior for  $\theta$  based on the default priors in Tibshirani (1989) or Datta and Ghosh (1995), or a fiducial distribution for  $\theta$  can be obtained based on the argument in Hannig (2013, Example 5.1). However, our efficiency principle tells us that, to construct a quality IM for  $\theta$ , we must reduce the dimension of the auxiliary variable as much as possible; this dimension reduction step will directly incorporate the marginal nature of the problem and yield results at least as efficient as Bayes and fiducial. See Martin and Liu (2015c, Sec. 4.3) for details.

## 7 Conclusion

In this paper, we have argued that the conversion of frequency probabilities coming from the posited sampling model into belief probabilities is the fundamental step in making inference, and we have provided a corresponding definition of valid, prior-free, probabilistic inference. Given this definition, it is now possible to assess the appropriateness of existing frameworks for inference relative to this definition. We concluded, in particular, that the frequentist and (objective) Bayesian frameworks are not appropriate. The key to the achieving probabilistic inference is a notion of a “predictable quantity,” something that is unobserved but has a known distribution. The auxiliary variable  $U$  in (5) is one example. Fiducial inference is based on a relationship like (5) but the validity condition (1) cannot be achieved in general, in part because the  $X$  and  $U$  are too tightly linked in that formulation. By treating the unobserved value of  $U$  as the target, the IM approach proposes to predict that unobserved value with a random set. The resulting IM output is a belief function, which we argued was natural for summarizing evidence. Furthermore, under a mild condition on the random set, the IM’s output is valid and, therefore, suitable for scientific inference. If desired, one can easily construct a decision procedure based on the IM output with guaranteed frequentist error rate control.

Stripped of the technical details of random sets, belief functions, etc, the IM approach is intuitively clear. The only potential obstacle in the formulation is the non-uniqueness

---

<sup>6</sup><http://stat.rutgers.edu/conferences/bff2016>



of the association in (5). Versions of the association with lower-dimensional auxiliary variables are preferable, which was the message from the *efficiency principle*, but it is not clear how to compare different versions having the same dimension.

Since the choice of random set affects the IM’s efficiency, a theory of “optimal” random sets is desirable. Steps in this direction have been made in the case of a fixed assertion  $A$ . However, as pointed out in Remark 2, if more than one assertion is of interest—even just  $A$  and  $A^c$ —then a gain of efficiency for one assertion result in a loss of efficiency for another. Using different random sets for each assertions is not the answer, since there may be concerns about interpretation (Martin and Liu 2014b, Sec. 4.2). So, the optimal random set must somehow balance efficiency across all the assertions of interest. Martin et al. (2016) makes some steps in this direction, but more work is needed.

We have focused here on the situation where no prior information is available, but one could argue that it is rare that *nothing* is known about the parameter of interest. In that case, it would make sense to incorporate that information into the analysis. The prior-free approach presented here can be extended by first recognizing that the IM’s inferential output  $b_x$  in (7) can be re-expressed as itself combined with the vacuous/ignorance prior belief in Section 4.1 according to, say, Dempster’s rule of combination (e.g., Shafer 1976). However, other more informative prior beliefs are possible—some that can even encode “partial” prior beliefs, as opposed to Bayesian priors that require the data analyst to specify beliefs about everything—and there may be other more appropriate rules by which the different forms of beliefs should be combined.

Many would say that surge of interest in “big-data” is a great opportunity for statisticians to have an impact. While this is certainly true, there is a lurking danger as well: in the excitement of tackling new problems, it may be easy to lose sight of what statistics is, what inference means, and the important role we have to play in the advancement of science. Barnard (1985) writes

*It seems to be useful for statisticians generally to engage in retrospection at this time, because there seems now to exist an opportunity for a convergence of view on the central core of our subject. Unless such an opportunity is taken there is a danger that the powerful central stream of development of our subject may break up into smaller and smaller rivulets which may run away and disappear into the sand.*

Providing valid probabilistic measures of uncertainty is essential for science in general and statistics in particular, so we hope that this paper will inspire others to embrace this point of view. With the sound logical framework described here, together with the ever-advancing computing technology, we believe that statisticians are poised to make important and timely contributions in today’s challenging big-data problems.

## Acknowledgments

The authors are grateful to Professor Harry Crane for his very helpful comments on an earlier draft of this paper.

## References

- Aldrich, J. (2000). Fisher’s “Inverse Probability” of 1930. *Int. Statist. Rev.*, 68(2):155–1172.
- Aldrich, J. (2008). R. A. Fisher on Bayes and Bayes’ theorem. *Bayesian Anal.*, 3(1):161–170.
- Barnard, G. A. (1985). A coherent view of statistical inference. Technical report, Department of Statistics and Actuarial Science, University of Waterloo.
- Barnard, G. A. (1995). Pivotal models and the fiducial argument. *Int. Statist. Rev.*, 63(3):309–323.
- Berger, J. (2006). The case for objective Bayesian analysis. *Bayesian Anal.*, 1(3):385–402.
- Berger, J. O., Bernardo, J. M., and Sun, D. (2009). The formal definition of reference priors. *Ann. Statist.*, 37(2):905–938.
- Berger, J. O., Liseo, B., and Wolpert, R. L. (1999). Integrated likelihood methods for eliminating nuisance parameters. *Statist. Sci.*, 14(1):1–28.
- Berger, J. O. and Wolpert, R. L. (1984). *The Likelihood Principle*. Institute of Mathematical Statistics Lecture Notes—Monograph Series, 6. Institute of Mathematical Statistics, Hayward, CA.
- Birnbaum, A. (1962). On the foundations of statistical inference. *J. Amer. Statist. Assoc.*, 57:269–326.
- Box, G. E. P. (1980). Sampling and Bayes’ inference in scientific modelling and robustness. *J. Roy. Statist. Soc. Ser. A*, 143(4):383–430. With discussion.
- Chiang, A. K. L. (2001). A simple general method for constructing confidence intervals for functions of variance components. *Technometrics*, 43(3):356–367.
- Choquet, G. (1953–1954). Theory of capacities. *Ann. Inst. Fourier, Grenoble*, 5:131–295 (1955).
- Cox, D. R. and Hinkley, D. V. (1974). *Theoretical statistics*. Chapman and Hall, London.
- Creasy, M. A. (1954). Symposium on interval estimation: Limits for the ratio of means. *J. Roy. Statist. Soc. Ser. B.*, 16:186–194.
- Datta, G. S. and Ghosh, J. K. (1995). On priors providing frequentist validity for Bayesian inference. *Biometrika*, 82(1):37–45.
- de Finetti, B. (1972). *Probability, Induction and Statistics*. John Wiley & Sons, London-New York-Sydney. Wiley Series in Probability and Mathematical Statistics.
- Dempster, A. P. (1963). Further examples of inconsistencies in the fiducial argument. *Ann. Math. Statist.*, 34:884–891.

- Dempster, A. P. (2008). The Dempster–Shafer calculus for statisticians. *Internat. J. Approx. Reason.*, 48(2):365–377.
- Dempster, A. P. (2014). Statistical inference from a Dempster–Shafer perspective. In Lin, X., Genest, C., Banks, D. L., Molenberghs, G., Scott, D. W., and Wang, J.-L., editors, *Past, Present, and Future of Statistical Science*, chapter 24. Chapman & Hall/CRC Press.
- E, L., Hannig, J., and Iyer, H. (2008). Fiducial intervals for variance components in an unbalanced two-component normal mixed linear model. *J. Amer. Statist. Assoc.*, 103(482):854–865.
- Edwards, A. W. F. (1997). What did Fisher mean by “inverse probability” in 1912–1922? *Statist. Sci.*, 12(3):177–184.
- Efron, B. (1998). R. A. Fisher in the 21st century. *Statist. Sci.*, 13(2):95–122.
- Efron, B. (2013a). Bayes’ theorem in the 21st century. *Science*, 340(6137):1177–1178.
- Efron, B. (2013b). Discussion: “Confidence distribution, the frequentist distribution estimator of a parameter: a review” [mr3047496]. *Int. Stat. Rev.*, 81(1):41–42.
- Ermini Leaf, D. and Liu, C. (2012). Inference about constrained parameters using the elastic belief method. *Internat. J. Approx. Reason.*, 53(5):709–727.
- Evans, M. (2013). What does the proof of Birnbaum’s theorem prove? *Electron. J. Stat.*, 7:2645–2655.
- Evans, M. (2015). *Measuring Statistical Evidence Using Relative Belief*. Monographs in Statistics and Applied Probability Series. Chapman & Hall/CRC Press.
- Fidler, F., Thomason, N., Cummings, G., Fineh, S., and Leeman, J. (2004). Editors can lead researchers to confidence intervals, but can’t make them think. *Psychol. Sci.*, 15:119–126.
- Fieller, E. C. (1954). Symposium on interval estimation: Some problems in interval estimation. *J. Roy. Statist. Soc. Ser. B.*, 16:175–185.
- Fisher, R. A. (1973). *Statistical Methods and Scientific Inference*. Hafner Press, New York, 3rd edition.
- Fraser, D. A. S. (1968). *The Structure of Inference*. John Wiley & Sons Inc., New York.
- Fraser, D. A. S. (2011). Is Bayes posterior just quick and dirty confidence? *Statist. Sci.*, 26(3):299–316.
- Fraser, D. A. S. (2014). Why does statistics have two theories? In Lin, X., Genest, C., Banks, D. L., Molenberghs, G., Scott, D. W., and Wang, J.-L., editors, *Past, Present, and Future of Statistical Science*, chapter 22. Chapman & Hall/CRC Press.
- Fraser, D. A. S., Reid, N., Marras, E., and Yi, G. Y. (2010). Default priors for Bayesian and frequentist inference. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 72(5):631–654.

- Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*. Chapman & Hall/CRC, Boca Raton, FL, second edition.
- Ghosh, J. K., Delampady, M., and Samanta, T. (2006). *An Introduction to Bayesian Analysis*. Springer, New York.
- Ghosh, M. (2011). Objective priors: An introduction for frequentists. *Statist. Sci.*, 26(2):187–202.
- Gleser, L. J. and Hwang, J. T. (1987). The nonexistence of  $100(1 - \alpha)\%$  confidence sets of finite expected diameter in errors-in-variables and related models. *Ann. Statist.*, 15(4):1351–1362.
- Hannig, J. (2009). On generalized fiducial inference. *Statist. Sinica*, 19(2):491–544.
- Hannig, J. (2013). Generalized fiducial inference via discretization. *Statist. Sinica*, 23(2):489–514.
- Hannig, J., Iyer, H., Lai, R. C. S., and Lee, T. C. M. (2015). Generalized fiducial inference: A review. *J. Amer. Statist. Assoc.*, to appear.
- Hannig, J., Iyer, H., and Patterson, P. (2006). Fiducial generalized confidence intervals. *J. Amer. Statist. Assoc.*, 101(473):254–269.
- Hannig, J. and Lee, T. C. M. (2009). Generalized fiducial inference for wavelet regression. *Biometrika*, 96(4):847–860.
- Kadane, J. B. (2011). *Principles of Uncertainty*. Texts in Statistical Science Series. CRC Press, Boca Raton, FL. <http://uncertainty.stat.cmu.edu>.
- Kass, R. E. and Wasserman, L. (1996). The selection of prior distributions by formal rules. *J. Amer. Statist. Assoc.*, 91(435):1343–1370.
- Lai, R. C. S., Hannig, J., and Lee, T. C. M. (2015). Generalized fiducial inference for ultrahigh dimensional regression. *J. Amer. Statist. Assoc.*, 110:760–772.
- Lehmann, E. L. and Casella, G. (1998). *Theory of Point Estimation*. Springer Texts in Statistics. Springer-Verlag, New York, second edition.
- Lindley, D. V. (2014). *Understanding Uncertainty*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, revised edition.
- Liu, D., Liu, R. Y., and Xie, M. (2015). Multivariate meta-analysis of heterogeneous studies using only summary statistics: efficiency and robustness. *J. Amer. Statist. Assoc.*, 110(509):326–340.
- Martin, R. and Liu, C. (2013). Inferential models: A framework for prior-free posterior probabilistic inference. *J. Amer. Statist. Assoc.*, 108(501):301–313.
- Martin, R. and Liu, C. (2014a). Discussion: Foundations of statistical inference, revisited. *Statist. Sci.*, 29:247–251.

- Martin, R. and Liu, C. (2014b). A note on  $p$ -values interpreted as plausibilities. *Statist. Sinica*, 24:1703–1716.
- Martin, R. and Liu, C. (2015a). Conditional inferential models: Combining information for prior-free probabilistic inference. *J. R. Stat. Soc. Ser. B*, 77(1):195–217.
- Martin, R. and Liu, C. (2015b). *Inferential Models: Reasoning with Uncertainty*. Monographs in Statistics and Applied Probability Series. Chapman & Hall/CRC Press.
- Martin, R. and Liu, C. (2015c). Marginal inferential models: Prior-free probabilistic inference on interest parameters. *J. Amer. Statist. Assoc.*, 110:1621–1631.
- Martin, R. and Walker, S. G. (2014). Asymptotically minimax empirical Bayes estimation of a sparse normal mean vector. *Electron. J. Stat.*, 8(2):2188–2206.
- Martin, R., Xu, H., Zhang, Z., and Liu, C. (2016). Valid uncertainty quantification about the model in a linear regression setting. Unpublished manuscript, [arXiv:1412.5139](#).
- Mayo, D. (2014). On the Birnbaum argument for the strong likelihood principle. *Statist. Sci.*, 29(2):227–239.
- McCullagh, P. (2002). What is a statistical model? *Ann. Statist.*, 30(5):1225–1310. With comments and a rejoinder by the author.
- McCullagh, P. and Nelder, J. A. (1983). *Generalized Linear Models*. Monographs on Statistics and Applied Probability. Chapman & Hall, London.
- Molchanov, I. (2005). *Theory of Random Sets*. Probability and Its Applications (New York). Springer-Verlag London Ltd., London.
- Nguyen, H. T. (1978). On random sets and belief functions. *J. Math. Anal. Appl.*, 65(3):531–542.
- Nguyen, H. T. (2006). *An Introduction to Random Sets*. Chapman & Hall/CRC, Boca Raton, FL.
- Nuzzo, R. (2014). Scientific method: Statistical errors. *Nature*, 506:150–152.
- Reid, N. and Cox, D. R. (2015). On some principles of statistical inference. *Int. Statist. Rev.*, 83(2):293–308.
- Robert, C. P. (2013). Discussion: “Confidence distribution, the frequentist distribution estimator of a parameter: a review” [mr3047496]. *Int. Stat. Rev.*, 81(1):52–56.
- Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Statist.*, 12(4):1151–1172.
- Savage, L. J. (1972). *The Foundations of Statistics*. Dover Publications, Inc., New York, revised edition.
- Schweder, T. and Hjort, N. L. (2002). Confidence and likelihood. *Scand. J. Statist.*, 29(2):309–332.

- Schweder, T. and Hjort, N. L. (2016). *Confidence, Likelihood, Probability: Statistical Inference with Confidence Distributions*. Cambridge Univ. Press.
- Shafer, G. (1976). *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, N.J.
- Shafer, G. (1979). Allocations of probability. *Ann. Probab.*, 7(5):827–839.
- Shafer, G. (2007). From Cournot’s principle to market efficiency. In Touffut, J.-P., editor, *Augustin Cournot: Modelling Economics*, pages 55–95. Edward Elgar.
- Shafer, G. and Vovk, V. (2006). The sources of Kolmogorov’s *Grundbegriffe*. *Statist. Sci.*, 21(1):70–98.
- Stein, C. (1959). An example of wide discrepancy between fiducial and confidence intervals. *Ann. Math. Statist.*, 30:877–880.
- Tibshirani, R. (1989). Noninformative priors for one parameter of many. *Biometrika*, 76(3):604–608.
- Trafimowa, D. and Marks, M. (2015). Editorial. *Basic Appl. Soc. Psych.*, 37(1):1–2.
- Wang, C. M., Hannig, J., and Iyer, H. K. (2012). Fiducial prediction intervals. *J. Statist. Plann. Inference*, 142(7):1980–1990.
- Wasserman, L. A. (1990). Belief functions and statistical inference. *Canad. J. Statist.*, 18(3):183–196.
- Weerahandi, S. (1993). Generalized confidence intervals. *J. Amer. Statist. Assoc.*, 88(423):899–905.
- Xie, M. and Singh, K. (2013). Confidence distribution, the frequentist distribution of a parameter – a review. *Int. Statist. Rev.*, 81(1):3–39.
- Xie, M., Singh, K., and Strawderman, W. E. (2011). Confidence distributions and a unifying framework for meta-analysis. *J. Amer. Statist. Assoc.*, 106(493):320–333.
- Yager, R. and Liu, L., editors (2008). *Classic Works of the Dempster–Shafer Theory of Belief Functions*, volume 219. Springer, Berlin.
- Zabell, S. (1989). R. A. Fisher on the history of inverse probability. *Statist. Sci.*, 4(3):247–263. With comments by Robin L. Plackett and G. A. Barnard and a rejoinder by the author.
- Zabell, S. L. (1992). R. A. Fisher and the fiducial argument. *Statist. Sci.*, 7(3):369–387.